Sparse Clustering of Noisy Signals in the Wavelet Domain

Tom Hope, Avishai Wagner, Or Zuk
Department of Statistics, Hebrew University of Jerusalem, Israel
{tom.hope,avishai.wagner,or.zuk}@mail.huji.ac.il

***Abstract*—We propose Sparse Clustering with Wavelets (SPARCWave) - a simple and efficient time-series clustering method particularly suited for low Signal-to-Noise Ratio (SNR) by a "built-in" shrinkage of wavelet coefficients based on their contribution to the clustering information. We also use group sparsity constraints to cluster multivariate signals.**

I. Background and previous work

Clustering of high-dimensional signals, sequences or functional data is a common task arising in many domains [1]. Clustering is often based on the pairwise distances between signals - but in the low Signal-to-Noise Ratio (SNR) scenario these distances may be unreliable. An apparent solution is to first smooth each signal, independently of the others – but, as we show, this may be sub-optimal as potentially important clustering information could be lost when noise is high.

Representing signals in the wavelet domain can be useful for clustering them. Giacofci et al. [2] first perform wavelets-based denoising for each signal, and then use model-based clustering on the union set of non-zero coefficients after shrinkage. However, informative features can be thresholded by single-signal denoising due to the lack of global information highlighting their importance (See Figure 1). Antoniadis [3] proposed extracting features representing the contribution of each wavelet scale $j \in \{1, \dots, J\}$ to the total energy of the signal. Here too, low SNR can lead to loss of informative features.

We combine the Discrete Wavelet Transform (DWT) and Sparse K-means [4], by formulating and optimizing the Sparse K-means objective in the wavelet domain. This leads to shrinkage which (i) uses "global" cross-signal information, and (ii) is geared towards preserving clustering information (rather than the signal for individual curves – see Figure 2). The method is shown to yield improved clustering performance, compared to methods available in the literature, in simulations on both univariate and multivariate signals.

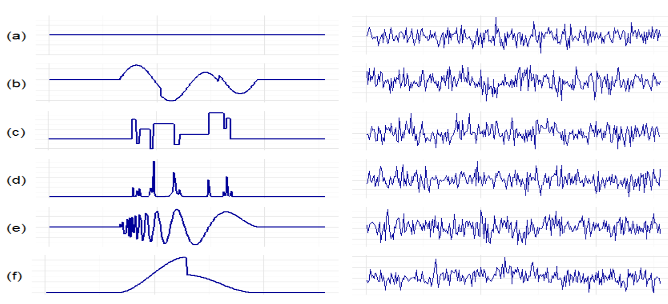

Fig. 1: **Simulated cluster data**. True cluster centers (Left) : (a) Flat curve, (b) Heavisine, (c) Blocks, (d) Bumps, (e) Doppler, (f) Piecewise polynomial, and actual noisy data sampled from each cluster with additive Gaussian noise (Right). SNR is too low to allow individual curve smoothing to reliably estimate cluster centers.

II. Methods

A. Sparse Clustering with Wavelets – Univariate Signals

Consider n instances (signal vectors) $\mathbf{x}^{(i)} \in \mathbb{R}^T$, and let $\mathbf{v}^{(i)}$ be their DWT transforms. Let $d_{i_1,i_2,j}$ be the squared (Euclidean) distance between $\mathbf{v}^{(i_1)}$ and $\mathbf{v}^{(i_2)}$ over coordinate j, $d_{i_1,i_2,j} \equiv (\mathbf{v}_j^{(i_1)} - \mathbf{v}_j^{(i_2)})^2$. Let C_k be the set of indices corresponding to cluster k with $|C_k| = n_k$. Finally, let $\mathbf{w}$ be a vector of weights, and s a tuning parameter bounding $\|\mathbf{w}\|_1$. We get the following Sparse K-means constrained optimization problem, where sparsity is promoted in the wavelet domain:

$$\underset{C_1,C_2,\dots C_k,\mathbf{w}}{\text{argmax}} \Big\{ \sum_{j=1}^{T} w_j \big(\frac{1}{n} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} d_{i_1,i_2,j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i_1,i_2 \in C_k} d_{i_1,i_2,j} \big) \Big\}$$
$$\text{subject to} \quad \|\mathbf{w}\|_2^2 \le 1, \|\mathbf{w}\|_1 \le s, w_j \ge 0 \quad \forall j = 1,..,n \tag{1}$$

The objective function is a decomposition of the weighted Between-Cluster-Sum-of-Squares (BCSS) into a sum over each coordinate j. For signals having good sparse approximation in the wavelet domain, we expect that clustering information will also be localized to a few wavelets coefficients, with high fitted weights w_j corresponding to informative coefficients (See Figure 2).

B. Tuning parameter selection

In [4], the parameter s is chosen with a permutation-based "Gap Statistic" method. This method may not be suitable when the number of instances is relatively low, as often occurs in the case of functional data, and tends to leave weights $\mathbf{w}$ arbitrarily small but non-zero [6]. We tried two additional approaches. First, we simply run the "Gap Statistic" method, and select the features corresponding to the top q^{th} percentile of weights $\mathbf{w}$ (e.g. $q = 99\%$). We also developed the following method to select s. Start with small $s_0 \ge 1$, solve eq. (1), and record the number of non-zero weights $\omega_0 = \|\mathbf{w}\|_0$. Gradually increase s by a fixed $\epsilon > 0$, $s_{i+1} \equiv s_i + \epsilon$ and repeat until s is large enough such that either $\|\mathbf{w}\|_0 = T$ (i.e. all features are selected) or the l_1 constraint $\|\mathbf{w}\|_1 \le s$ is guaranteed to be satisfied ($s = \sqrt{T}$) yielding an array $\{\omega_0, \omega_1, \dots\}$. Compute the ratios $\frac{\omega_{i+1}}{\omega_i}$, find $i_{\max} = \underset{i}{\text{argmax}} \frac{\omega_{i+1}}{\omega_i}$, and select $s = s_{i_{\max}}$. In words, we watch for the "largest jump" in the number of selected features, and take the largest s before this jump. Finally, in both approaches we use the selected features in standard K-means, ignoring their weights – we select features with Sparse K-means, and then use them for "unbiased" clustering. Both our methods resulted in improved accuracy in simulations.

C. Clustering Multivariate Signals with Group Sparsity

In many cases, each instance is actually a multivariate signal $\mathbf{X}^{(i)} \in \mathbb{R}^{G \times T}$ where G is the number of variables and T is as above (See Figure 4). We use a direct generalization of our method for univariate signals. We transform each row of $\mathbf{X}^{(i)}$ with DWT to get $\mathbf{V}^{(i)}$, and use $d_{i_1,i_2,j} \equiv \sum_{g=1}^{G} (\mathbf{V}_{g,j}^{(i_1)} - \mathbf{V}_{g,j}^{(i_1)})^2$. These $d_{i_1,i_2,j}$ values are used in solving (1) as in the single-curve case, yielding a distance which is still decomposable over j. This problem formulation (termed SPARCWave Group) yields group sparsity penalties as a by-product: It is equivalent to concatenating the rows of each $\mathbf{V}^{(i)}$ into one vector of length $G \times T$ and solving a univariate problem with the additional constraints that $w_i = w_{i+T} = \ldots = w_{i+GT}$. We also applied the concatenation approach without group sparsity (termed SPARCWave Concat), yet group sparsity improved results (See Figure 5).

III. Simulation Results

For univariate signals, we take curves of dimension 256 from [5] (using the *waveband* R package) and pad them with 128 zeros on both ends resulting in 6 cluster centers with $T = 512$ (see Figure 1). We apply additive Gaussian noise $N(0, \sigma^2)$ to generate individual signals. Signals are clustered using K-means (picking the best of 100 starts), Sparse K-means on the raw data as in [4], and the methods of [2] (using the *curvclust* R package) and [3] (See Figure 3). We use the Adjusted Rand Index as a measure of clustering accuracy [7].

For multivariate signals, we select univariate curves as depicted in Figure 4, with each univariate signal of dimension 128 padded with 64 zeros on each end, resulting in 5 cluster centers with $G = 3, T = 256$. We compare our method to K-means and Sparse K-means on the concatenated data, to a Hidden Markov Model approach in [8] and use the multivariate PCA-similarity measure [9] to construct a pairwise distance matrix, which we use in spectral clustering [10].

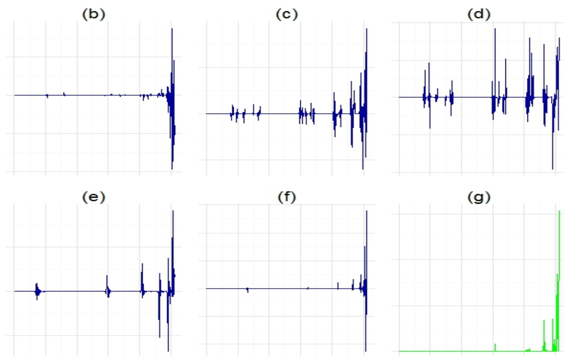

Fig. 2: **SPARCWave selected wavelet coefficients**. Wavelet coefficients for the univariate curves from Figure 1 (excluding the trivial flat curve) in blue, and the SPARCWave clustering weights (with noise level $\sigma = 2.5$) in green. Wavelet coefficients in each curve are ordered from the finest (left) to the coarsest level (right). Although a few wavelet coefficients at fine resolution are large for some of the curves, the informative coefficients for clustering are all at the coarsest levels.

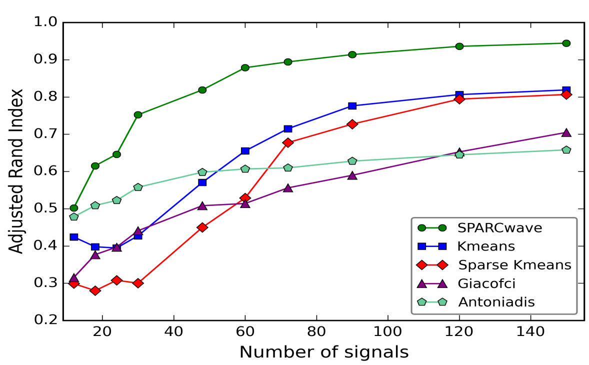

Fig. 3: **Univariate simulation results**. Average Adjusted Rand Index as a function of number of signals over 2000 simulations, for the simulation described in text with $\sigma = 2.75$ and clusters of equal size. Both K-means, which does not exploit sparsity, and Sparse K-means in the time domain, show inferior accuracy compared to SPARCWave.

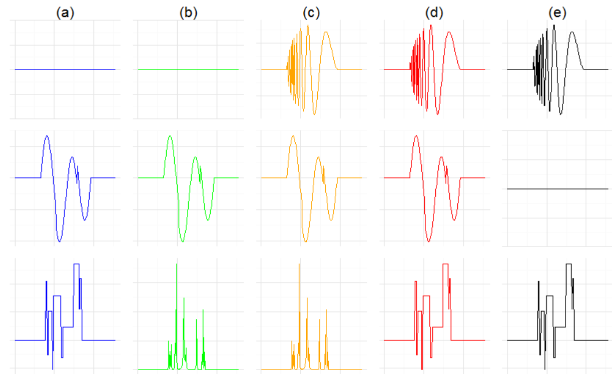

Fig. 4: **Multivariate simulations**. Each column (color) is a multivariate true cluster center. We add white noise to each center, forming our simulated data. Some pairs of clusters (e.g. (c) and (d)) are identical in two of the three curves, and can be distinguished only based on the third. Each group in the group sparsity problem formulation in the text is formed by taking the same wavelet coefficient in all three curves.

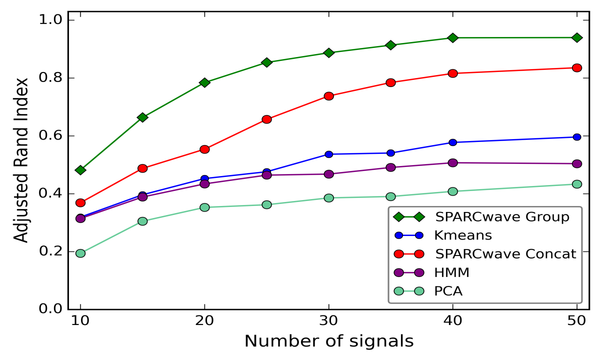

Fig. 5: **Multivariate simulation results**. The average Adjusted Rand Index is shown as function of number of signals over 2000 simulations for different methods, for the simulation described in text with $\sigma = 1.75$ and clusters of equal size. Concatenating the three curves into a single curve and enforcing sparsity in the wavelet domain (SPARCWave Concat) improves upon K-means. However, this method ignores the multivariate nature of the curves, and exploiting this information using group sparsity (SPARCWave Group) yields the most accurate clustering.

References

[1] T. W. Liao, Clustering of Time Series Data: a Survey, *Pattern Recognition*, 2005.

[2] M. Giacofci, S.L. Lacroix, G. Marot and F. Picard, Wavelet-Based Clustering for Mixed-Effects Functional Models in High Dimension, *Biometrics*, 2013.

[3] A. Antoniadis, X. Brossat and J. Cugliari, Clustering Functional Data Using Wavelets, *International Journal of Wavelets, Multiresolution and Information Processing* , 2011.

[4] D. Witten and R. Tibshirani, A Framework for Feature Selection in Clustering, *Journal of the American Statistical Association*, 2010.

[5] D. L. Donoho, De-noising by Soft-Thresholding, *IEEE Transactions on Information Theory*, 1994.

[6] M. Azizyan, A. Singh and Wei Wu., Experimental Evaluation of Feature Selection Methods for Clustering. https://www.ml.cmu.edu/research/dap-papers/dap_azizyan.pdf

[7] S. Wagner and D. Wagner, Comparing Clusterings: an Overview, *Karlsruhe: Universitt Karlsruhe, Fakultt fr Informatik*, 2007.

[8] S. Ghassempour, F. Girosi and A. Maeder, Clustering Multivariate Time Series Using Hidden Markov Models, *Int. J. Environ. Res. Public Health*, 2014.

[9] K. Yang and C. Shahabi, A PCA-based Similarity Measure for Multivariate Time Series, *Proceedings of the 2nd ACM international workshop on Multimedia databases*, 2004.

[10] A. Y. Ng, M. I. Jordan and Y. Weiss, On Spectral Clustering:Analysis and an Algorithm, *Advances in Neural Information Processing System (NIPS)*, 2002.